GENERALIZED DRIVEN DECODING FOR SPEECH RECOGNITION SYSTEM COMBINATION

Benjamin Lecouteux, Georges Linares, Yannick Estève, Guillaume Gravier

GENERALIZED DRIVEN DECODING FOR SPEECH RECOGNITION SYSTEM COMBINATION

Benjamin Lecouteux (1), Georges Linarès (1), Yannick Estève (2), Guillaume Gravier (3)
LIA, Avignon (France) (1), LIUM, Le Mans (France) (2), IRISA, Rennes (France) (3)

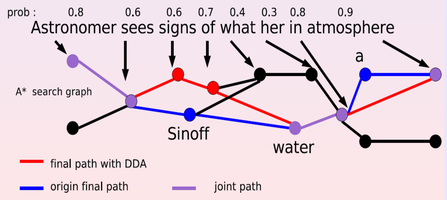
The Driven Decoding Algorithm

Principle of driven decoding algorithm (DDA)

The DDA based combination :

- A first recognition pass using an auxiliary ASR system
  - Auxiliary system provides the one-best hypothesis h_{aux}
  - The auxiliary transcript drives the main search algorithm
- DDA is an **integrated approach** for system combination

- Trancripts drive A* decoding :
  - A* search is synchronised to the transcript
  - Linguistic probabilities are dynamically rescored
- Rescoring is based on posteriors



Anatomy of the Speeral decoder

- Large vocabulary continuous speech recognition system
- HMM-based acoustic modeling
- Trigram language models
- Search : derived from a A* search algorithm operating on a lattice of phonemes
- Exploration is supervised by the function $F(h_n)$ evaluating the probability of h_n crossing the node n:

$$F(h_n) = g(h_n) + p(h_n) \quad (1)$$

DDA step 1 : on-demand synchronization

- Speeral speech recognition system generates hypotheses as the phoneme-lattice is explored
- A* is an asyncronous decoder :
  - Hypotheses are extended or left according to $F()$
  - Leave a path leads to backtracking
- DDA synchronizes the current hypothesis and the auxiliary transcript
- Synchronization by fast DTW algorithm

DDA step-2 : transcript to hypothesis matching score

- Linguistic probabilities are modified using the following rescoring rule:

$$L(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\beta}.\alpha(w_i)^{\beta} \quad (2)$$

$L(w_i|w_{i-2},w_{i-1})$ is the resulting linguistic score , $P(w_i|w_{i-2},w_{i-1})$ the initial probability, β an empirical fudge factor and $\alpha(w_i)$ is the confidence score of w_i given by :

$$\text{if } \theta(w_i) > 0 \text{ then } \alpha(w_i) = \phi(w_i)\ .\ \frac{\theta(w_i)}{\gamma} \text{ and } \beta = 0.6$$
$$\text{else } \beta = 0 \quad .$$

γ is the analysis window size reported by the edit distance ($\gamma = 4$) and $\phi(w_i)$ posteriors from word w_i of the auxiliary system.

Experimental framework

The LIA system

- System involved in the ESTER evaluation campaign :
  - Speeral decoder
  - Alize-based segmenter
  - 65K lexicon; 20M trigrams estimated on about 200M of words
  - 2 decoding pass (MLLR adaptation)
  - 8xRT on a standard desktop computer

The LIUM system

- Based on the CMU Sphinx 3.3 decoder (beam search algorithm)
- 4-gram word-lattice rescoring process
- Context-depedent acoustic models trained on Ester materials
- SAT-based adaptation
- The entire process runs under 12xRT

The IRISA system

- Based on word-synchronous beam-search algorithm
- HMM acoustic modeling and n-gram linguistic models with a vocabulary of 64k words
- The system operates in four steps
- (1) Context-independent acoustic models with a trigram LM
- (2) The graph is rescored with a 4gram LM and context-dependent models
- (3) MLLR speaker adaptation
- (4) Consensus decoding is applied to the 1000-best sentence hypotheses

- ASR systems are assessed on 3 hours of radio broadcast from ESTER corpus
- The Driven Driven Algorithm (DDA) is used here during the second pass

Baseline results

	F. Inter	F. Info	RFI
LIA	21.1	22.2	24.6
LIUM	18.5	18.9	25.6
IRISA	21.4	21.8	25.6
DDA-IRISA-P1	19.6	19.3	23.5
DDA-IRISA-P2	18.7	18.7	22.2
DDA-LIUM-P1	17.8	18.1	22.4
DDA-LIUM-P2	17.2	17.8	21.5

Table: Word error rates for DDA combination of Speeral with an LIUM system (*DDA-LIUM*) and IRISA system (*DDA-IRISA*) with (*P1* and without (*P2*) unsupervised speaker adaptation. Experiments performed of 3 hours of French broadcast news from the ESTER corpus.

Confusion network driven decoding

Principle :

- As with the single best output, the combination method operates at the search level
- The current word utterance is dynamically mapped to the confusion network.
- The alignment step allows to extract the best projection of the hypothesis in the network
- The rescoring problem is similar to the one-best driven decoding case
- Performance are very close to the one obtained with the more simple one-best driven decoding

	F. Inter	F. Info	RFI
LIUM	18.5	18.9	25.6
DDA-LIUM-P1	17.8	18.1	22.4
DDA-LIUM-P2	17.2	17.8	21.5
DDA-WCN-LIUM-P1	17.7	18.1	22.3
DDA-WCN-LIUM-P2	17.2	17.8	21.5

Table: Word error rates for confusion network driven decoding (*DDA-WCN*), according to the decoding pass. Results are compared to the ones of the best single system (LIUM) and to the best one-best DDA system (*DDA-LIUM*)

Multi system combination

Two-Level ROVER-DDA combination

- Relies on a first merging step where all auxiliary transcripts are merged
- We use ROVER for merging LIUM and IRISA system outputs
- The word confidence scores of the output are computed by averaging the confidence scores of words in each single system output
- The resulting transcript is then used as an auxiliary hypothesis

Integrated DDA-based combination

- All auxiliary systems outputs are submitted
- For each of them, a matching score is computed according to independent transcript-to-hypothesis synchronization
- all linguistic scores are merged by the log-linear combination extended to n systems:

$$L(w_i|w_{i-2},w_{i-1}) = P(w_i|w_{i-2},w_{i-1})^{1-\beta}.\frac{1}{N}\sum_{k=0}^{N}\alpha_k(w_i)^{\beta_k} \quad (3)$$

where β is the averaged β_k as defined in equation 2, α_k are the posteriors provided by the system k and N the number of auxiliary systems.

Results

	F. Inter	F. Info	RFI
LIUM	18.5	18.9	25.6
ROVER-3	17.1	18.2	22.5
2-Level DDA-ROVER	16.8	17.3	21.3
DDA-3	16.7	17.0	20.6
DDA-3+ROVER	16.0	16.4	20.7

Table: Word error rates of multiple-system combination according to the combination schemes : the baseline ROVER combination of the 3 single systems (*ROVER-3*), the 2-level method (*2-Level DDA-ROVER*), the full DDA-integration (DDA-3) of auxiliary systems, and the ROVER combination of all systems including DDA-3 (DDA-3+ROVER). This last one obtains the best results with a WER decrease of about 15.7% relative with respect to the best single system (LIUM).

	F. Inter	F. Info	RFI
Best single system(LIUM)	18.5	18.9	25.6
DDA-3	16.7	17.0	20.6
ORACLE-3	10.3	10.5	14.5
DDA-3+ROVER	16.0	16.4	20.7
ORACLE DDA+ROVER	9.8	10.0	13.6

Table: Analysis of DDA by comparison to ROVER and Oracle measures.

Driven decoding analysis and Conclusion :

- Linguistic rescoring allows to guide the search toward alternative paths
- DDA may not be considered as an on-line vote method
- DDA is an integrated approach allowing a new exploration of the search graph
- DDA outperforms the ROVER approach
- ROVER combination of DDA-3 and all single systems outperfoms the pure DDA approach

- WCN-driven decoding improves the primary system but performance are very close to the one obtained with the more simple one-best driven decoding
- By using DDA-based cross-site system combination and a final ROVER pass, we obtained a global absolute gain of about 3.3% WER (15.7% relative gain).

This research is supported by the ANR (Agence Nationale de la Recherche), EPAC project ANR-06-MDCA-006